Overview

Deriving scientific information from large, complex datasets can motivate large, complex statistical models, in fields as diverse as astronomy [Regier et al., 2019], economics [Meager, 2019], phylogenetics [Pritchard et al., 2000], and many more. As models grow in complexity, the need to interrogate their assumptions and to propagate uncertainty amongst their components grows, as does the computational cost of doing so using traditional statistical methods. Many classical procedures designed to address these concerns, such as Markov Chain Monte Carlo (MCMC), cross validation, or re-estimating a model under a range of modeling assumptions, can be prohibitively expensive in many modern problems.

To address this gap, my work employs *sensitivity analysis*, applied not merely in the traditional sense of assessing the risk of to imprecise modeling assumptions (though I do pursue this traditional role as well), but also to quantify freqeuntist sampling properties and propagate uncertainty in Bayesian procedures. Though conceptually unified, my work is diverse in applications, touching many of the core activities of modern data analysis, from cross-validation [Giordano et al., 2019b,a], scalable Bayesian posterior inference [Giordano et al., 2018a,b], the bootstrap [Giordano and Broderick, 2020], and more. A recurrent theme of my work is adapting classical *theoretical tools* [Reeds, 1976, Gustafson, 1996] to *modern computing environments* equipped with scalable, general purpose automatic differentiation software [Baydin et al., 2017, Carpenter et al., 2015].

Data sensitivity: cross validation and frequentist variance

Frequentist variability is ultimately concerned with the outcome of an estimation procedure if the data were drawn from the same distribution as but different from that observed. Similarly, all forms of cross-validation (CV) evaluates a statistic if parts of the observed data had been ablated. Both of these procedures can be treated by sensitivity analysis, where sensitivity is to the dataset itself.

Accuracy bounds for approximate cross validation. To perform leave-one-out CV (LOO-CV), one re-runs an estimation procedure with each datapoint left out. In full, LOO-CV requires as many re-runs procedures as there are datapoints, and each re-run is expected to be quite close to the original fit. Rather than re-running exactly, one can use a Taylor series to approximate the effect of removing a single data point; since the dataset with one point left out is, in some sense, "close" to the original dataset, the Taylor series can be expected to perform well.

Prior to our work, this idea had been suggested both in the machine learning literature [Rad and Maleki, 2018, Koh and Liang, 2017] as well as in the classical statistical literature under the name "infinitesimal jackknife" [Jaeckel, 1972, Shao and Tu, 2012]. However, the ML work appeared unaware of the statisical precedent, and both treatments required unrealistic theoretical conditions for the accuracy of the Taylor series: specifically, that the gradients of the objective function be uniformly bounded, a condition that is rarely satisfied in scientific practice, even in the simplest possible example of using maximum likelihood to estimate the sample mean of a normal distribution.

In Giordano et al. [2019b], we provide a more realistic set complexity condition under which the Taylor series is accurate, eschewing the need for bounded gradients, and synthesizing the classical statistics and ML literatures. Also unlike previous work, our theory was purely finite sample, implying the asymptotica results of prior work as a corollary. We demonstrated the accuracy of the technique on an unsupervised clustering problem from genomics [Shoemaker et al., 2015].

Sensitivity to removal of a small fraction of the data. Classical frequentist standard errors estimate the variability in an estimator that would result from the rarified thought experiment of re-sampling datsets from the same distribution that gave rise to the observed data. In the social sciences, this rarefied experiment rarely closely corresponds to reality, and one might be concerned if substantive conclusions could be overtuned by other minor perturbations to the data. For example, if a top-line conclusion of a study of the efficacy of cash transfers in a particular country [Angelucci and De Giorgi, 2009] could be reversed by removing a small percentage (say, 0.1%) of a dataset, one might hesitate to generalize one's conclusions to other countries, even if the result was "statistically significant" according to classical frequentist standard errors.

In Giordano et al. [2020], we address this fundamental question, extending our earlier results in Giordano et al. [2019b]. We find that problems with small signal-to-noise ratio but large datasets will be particularly non-robust to the removal of a small proportion of the data. Such a situation that obtains commonly in econometrics, and we find that the sign and statistical significance of estimated effects in a number of large, prominent econometric studies can be overturned by dropping only a small number of datapoints [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012]. Our robustness metric can be computed easily and automatically for any Z-estimator; we provide software and tractable finite-sample accuracy bounds for ordinary least squares and instrumental variables regression. More broadly, our work points to the importance of considering "practical significance" of effects in the social sciences rather than mere statistical significance.

Frequentist properties of Bayesian posteriors Bayesian measures is a powerful tool for coherently treating uncertainty in complex problems, but, when the model is misspecified, the estimated posterior uncertainty may not be meaningful. One can, however, always compute the frequentist sampling variability of a Bayesian posterior quantity, and such a quantity always remains meaningful, even if conceptually distinct from a posterior uncertainty Waddell et al. [2002], Kleijn and van der Vaart [2006], Huggins and Miller [2019].

By combining the frequentist IJ [Jaeckel, 1972, Shao and Tu, 2012, Giordano et al., 2019b] approach to frequentist variance with the MCMC-based measures of sensitivity [Gustafson, 2000, Giordano et al., 2018a], we are able to derive the Bayesian infinitesimal jackknife (IJ), which can be used to compute the frequentist variability of Bayeisan posterior means without bootstrapping or computing a maximum a-posteriori (MAP) estimate. In a work in progress, we extend the Bayesian central limit theorem of Lehmann and Casella [2006], Kass et al. [1990] to prove the consistency of the Bayesian IJ and show its accuracy as an approximation to the bootstrap for a larger number of examples, effectively allowing estimation of frequentist covariances orders of magnitude faster than the bootstrap. We demonstrate the accuracy of our method on datasets from election modeling [Gelman and Heidemanns, 2020], ecology [Kéry and Schaub, 2011], and most of the models from [Gelman and Hill, 2006, Stan Team, 2017].

Propagation of uncertainty in scalable Bayesian inference

One popular technique to scale Bayesian inference to massive problems is mean field variational Bayes (MFVB) [Wainwright and Jordan, 2008, Blei et al., 2017, Regier et al., 2019]. However, MFVB provides notoriously innacurate posterior uncertainty estimates, even in situations when it estimates the posterior means accurately. In [Giordano et al., 2018a], we develop a method to recover accurate posterior uncertainties from MFVB approximations without needing to fit a more complex model, or indeed to re-fit the original model. Computing the LRVB covaraince requires solving a linear system, which in scientific applications is often sparse and can be solved using iterative techniques such as conjugate gradient [Nocedal and Wright, 2006, Chapter 5]. We compare LRVB covariances to MCMC on a large number of real-world datasets, including logistic regression on an internet advertising dataset [Criteo Labs, 2014], the Cormack-Jolly-Seber model from ecology [Kéry and Schaub, 2011], and hierarchical generalized linear models from the social sciences [Gelman and Hill, 2006], demonstrating accurate posterior covariances computed over an order of magnitude faster than MCMC.

Prior sensitivity in Bayesian analysis

Bayesian techniques allows analysts to reason coherently about unknown parameters, but only if the user specifies a complete generating process for the parameters and data, including both prior distributions for the parameters and precise likelihoods for the data. Often, aspects of this model are at best a considered simplifcation, and at worst chosen only for computational convenience. It is critical to ask whether the analysis would have changed substantively had different modeling choices been made.

Bayesian nonparametrics. A commonly question in unsupervised clustering is how many distinct clusters are present in a dataset. Discrete Bayesian nonparametrics (BNP) allows the answer to be inferred using Bayesian inference, but one must specify a prior on how distinct clusters are generated [Ghosh and Ramamoorthi, 2003, Gershman and Blei, 2012]. A particularly common modeling choice is the stick-breaking representation of a Dirichlet process prior [Sethuraman, 1994], a mathematical abstraction which is arguably better justified by its computational convenience than its realism. Our workshop paper, Giordano et al. [2018b], fits a BNP model with variational Bayes [Blei and Jordan, 2006] using the standard, computationally convenient stick-breaking prior, but then uses sensitivity analysis to allow the user to explore alternative functional forms an order of magnitude faster than would be possible with refitting. In work currently in progress, we apply our method to a human genome dataset in phylogenetics taken from [Huang et al., 2011], and find that our method accurately discovers real excess prior sensitivity in a BNP version of the model `fastSTRUCTURE` [Raj et al., 2014].

Partial pooling in meta-analysis. A popular form of meta-analysis in econometrics is to place a hierarchial model on a set of related experimental results, which both "shrinks" the individual estimates towards a common mean, potentially decreasing mean squared error, and allowing direct estimation of the average effect and diversity of effects [Rubin, 1981, Gelman and Rubin, 1992]. These advantages come at the cost of positing a precise generative process for the effects in question, and it is reasonable to interrogate whether the estimation procedure is robust to varaibility in these effects. In Giordano et al. [2016], we apply sensitivity analysis to a published meta-analysis of the effectiveness of microcredit interventions in seven developing countries [Meager, 2019]. We find that the conclusion are highly sensitive to the assumed covariance structure between the base level of business profitability and the microcredit effect, a covariance which is *a priori* difficult to ascertain, automatically diagnosing an important source of epistemic uncertainty not captured by the Bayesian posterior.

Hyperparameter sensitivity for MCMC. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, and modern black-box MCMC tools such as `Stan` [Stan Development Team, 2020, Carpenter et al., 2017]. However, MCMC still often takes a long time to run, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models. A classical result from Bayesian robustness states that the sensitivity of a posterior expectation is given by a particular posterior covariance [Gustafson, 1996, Basu et al., 1996], though the result has not been widely used, arguably due in part to the lack of an automatic implementation. In my software package, Giordano [2020], I take advantage of the automatic differentiation capacities of `Stan` to provide automatic hyperparameter sensitivity for generic Stan models. In examples in the package `git` repository, I demonstrate the efficacy of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

References

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–153.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164*.

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the us elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Gershman, S. and Blei, D. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer Science & Business Media.

Giordano, R. (2020). Stansensitivity: automated hyperparameter sensitivity for Stan models.

Giordano, R. and Broderick, T. (2020). *The Bayesian Infinitesimal Jackknife for Variance.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., Meager, R., Huggins, J., and Jordan, M. I. (2016). Fast robustness quantification with variational Bayes. *arXiv preprint arXiv:1606.07153*.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2018b). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *arXiv preprint arXiv:1810.06587*.

Giordano, R., Meager, R., and Broderick, T. (2020). *An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?*

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Huang, L., Jakobsson, M., Pemberton, T., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J., Tishkoff, S., and Rosenberg, N. (2011). Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Jaeckel, L. (1972). The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ.

Kass, R., Tierney, L., and Kadane, J. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*.

Lehmann, E. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Rad, K. and Maleki, A. (2018). A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv Preprint*.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Reeds, J. (1976). *Ill (1976). On the definition of von Mises functionals*. PhD thesis, Ph. D. Thesis, Statistics, Harvard University.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through bayesian inference in julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Rubin, D. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6(4):377–401.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Team (2017). Stan example models wiki. Referenced on June 5th, 2020.

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ml and bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wainwright, M. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Machine Learning*, 1(1-2):1–305.